Original Article

Stress Testing Pathology Models with Generated Artifacts

Nicholas Chandler Wang[1], Jeremy Kaplan[1], Joonsang Lee[1], Jeffrey Hodgin[2], Aaron Udager[2], Arvind Rao[1]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA, [2]Department of Pathology, University of Michigan Medical School, Ann Arbor, MI, USA

Submitted: 20-Jan-2021 Revised: 23-Jun-2021 Accepted: 06-Jul-2021 Published: ***

Abstract

Background: Machine learning models provide significant opportunities for improvement in health care, but their "black-box" nature poses many risks. **Methods:** We built a custom Python module as part of a framework for generating artifacts that are meant to be tunable and describable to allow for future testing needs. We conducted an analysis of a previously published digital pathology classification model and an internally developed kidney tissue segmentation model, utilizing a variety of generated artifacts including testing their effects. The artifacts simulated were bubbles, tissue folds, uneven illumination, marker lines, uneven sectioning, altered staining, and tissue tears. **Results:** We found that there is some performance degradation on the tiles with artifacts, particularly with altered stains but also with marker lines, tissue folds, and uneven sectioning. We also found that the response of deep learning models to artifacts could be nonlinear. **Conclusions:** Generated artifacts can provide a useful tool for testing and building trust in machine learning models by understanding where these models might fail.

Keywords: Artifact, digital pathology, failure mode, machine learning, neural network, robustness

Introduction

As machine learning models increase in presence throughout health care, the need for tools to robustly measure their performance characteristics greatly increases. While the medical field has a long history of evaluating the performance of diagnostic tests, machine learning models pose new challenges to interpreting performance characteristics.[1] On the surface, both types of tests can digest complex biology to a singular result, but the potential sources of error involved in each dataset vary greatly. Where traditional diagnostic tests involve measurement of one or more analytes, machine learning models can interpret orders of magnitude more features from a rich dataset.[2] Just as understanding which other substances might interfere with a single analyte test is important, understanding what aspects of normal operation may impact a machine learning model is critical to evaluating its performance.[3,4] This work studies how to simulate some known artifacts in pathology and evaluates how they affect existing machine learning models.

A significant concern with applying machine learning models in health care is the possibility of accumulating undesired or biased errors when a model is used in a real-world setting.[5-7] Algorithms can fail by not taking into account known problems and could perpetuate inequities in health care. Obermeyer *et al*. found that a risk evaluation algorithm was biased against black patients because it used money spent on health care as a proxy for health.[8] Black patients at a given risk score were far more sick than their white counterparts, so because black patients have less access to health care, the algorithm recommended fewer health-care interventions for them.

Digital pathology models are some of the closest deep learning models to being used in a clinical setting. These models leverage advances in computer vision models to learn the process a pathologist might take in evaluating tissue morphologies and staining patterns to augment a pathologist's diagnostic evaluation. Several such products are reportedly under clinical development,[9] with several already gaining clinical approval in the European Union to help pathologists focus on regions of interest when evaluating prostate biopsies.[10,11]

We sought to probe two such models developed for research purposes, to understand their limitations. For the first model,

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_6_21

Address for correspondence: Dr. Arvind Rao, 100 Washtenaw Ave, Room 2305, Ann Arbor, MI 48109, USA.
E-mail: ukarvind@med.umich.edu

How to cite this article: Wang NC, Kaplan J, Lee J, Hodgin J, Udager A, Rao A. Stress testing pathology models with generated artifacts. J Pathol Inform 2021;12:54.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2021/12/1/54/333710

Coudray *et al*.[12] published a convolutional neural network model for classifying histological subtypes of nonsmall cell lung cancer, based on the InceptionV3 architecture.[13] Histological classification is vitally important in the proper diagnosis of all cancers, as it is a large determinant of what types of therapies are available for a patient. For example, most targeted therapies in lung cancer are only for adenocarcinomas.[14]

The second model and dataset we tested was an internal dataset of kidney histopathology samples. Instead of classification, the purpose of this convolutional neural network was to label the different tissue components present on a tissue sample. This segmentation task was meant to differentiate different components of tissue and label kidney glomeruli that were globally sclerosed and normal, in addition to labeling the other components of tissue that were present.

The enumeration, identification, and mitigation of artifacts are frequently discussed topics in the digital pathology space. Manufacturers have published white papers and protocol guidance to help pathologists be consistent in tissue staining and fixation, while academic pathologists have focused on reporting and understanding the causes of artifacts.[15-18] In addition, the challenges of whole-slide imaging, and best practices for creating a pipeline that can robustly handle the mixed quality of tissue samples, have been described by other groups.[19] Automated toolkits, such as HistoQC, have been created to identify slides or regions of slides that are affected by artifacts as a QC process.[20] In this study, we describe a tool for analyzing the performance of histopathology models by introducing artifacts and studying their effect.

Methods

Software description

The artifact generation package was developed as a custom-written Python module to generate synthetic artifacts. The goal of this module was to artificially create histology artifacts to evaluate the error bounds of machine learning systems when presented with failure states or poor-quality data. Histology is a well-studied field, and there is plenty of literature describing different types of artifacts, including common problems with histology samples. Seven different types of artifacts were simulated, and each was parameterized to allow for quantification of noise levels [Figure 1 and described in section 2.2].

A few general principles guided the development of this artifact generation toolkit. First, an artifact should be generated according to describable and explainable methodologies. However, to expand the variety of outputs, the specific parameters of an individual instance of an artifact should be driven by seeded random generation, as to be diverse and reproducible. The method should be able to run with the only required input being the image tile itself to reduce the overhead of the method and enable simple parallelization. Finally, each method should have user-configurable parameters to tune the artifacts in future projects. Together, this philosophy drove how these artifacts were generated, and could provide a framework for other contributors to add new artifacts.

Artifact generation methods

Bubbles

Bubbles create an artifact on top of the tissue on the slide and can be formed in a few ways. Air bubbles can be introduced by air getting underneath the coverslip, while nuclear bubbles can be caused by heat or other conditions that cause protein coagulation.[15,16] Our toolkit generates bubbles by creating a partially transparent layer to overlay on top of the tissue tile image. The locations of the bubbles are generated randomly across the image, and each of the bubbles is generated by a two-dimensional Gaussian distribution. The size, orientation, and width of these distributions are determined by their randomly generated covariance matrices, to create reasonable but small bubbles. A fixed number of these Gaussian distributions are summed together to create the set of bubbles. The edges of these bubbles are found and are colored a darkened version of the mean color of the image. The interior of the bubbles is colored a fixed light gray color. In future work, any of these colors or distributional settings could be varied and adjusted using the function input parameters.

Tissue fold

Tissue folds occur when the thin tissue slice is not uniformly spread out and folds back over itself, often due to heterogeneous material properties within the sample.[16] Toolkits exist for identifying these regions in a sample as part of an imaging quality control process.[21] We generated fold artifacts using a path along a randomized three-point spline, anchored on the edges of the images and containing a single intermediate knot. The spline region was then randomly shifted to another region of the image, with some warping to get a sample of tissue. The image was tiled by mirroring at the edges to allow for sampling beyond the edge of the original image.

The sampled tissue region was then overlaid on top of the original image, with a multiplicative combination after a slight Gaussian blur. This created a tissue fold of a single layer on top of the image. This process is repeated recursively over the same spline region to build up a fold of multiple layers. The default number of folded layers is two on top of the original image, as one layer would be sandwiched in the wrong direction, and another would be heading back in the correct orientation. This tissue sampling makes the assumption that locally, the texture and staining at the tile level are sufficiently consistent to make tissue folds.

Illumination

Uneven illumination artifacts can occur when the light source behind the histopathology slide is not consistent across the image. Existing methods can correct this illumination issue as a quality control step.[22] This illumination artifact is generated by adding together three wide Gaussian distributions to create an illumination map. The minimum and maximum illumination change is randomly generated within a narrow range of 80%–110% and used to rescale the Gaussian map. This rescaled

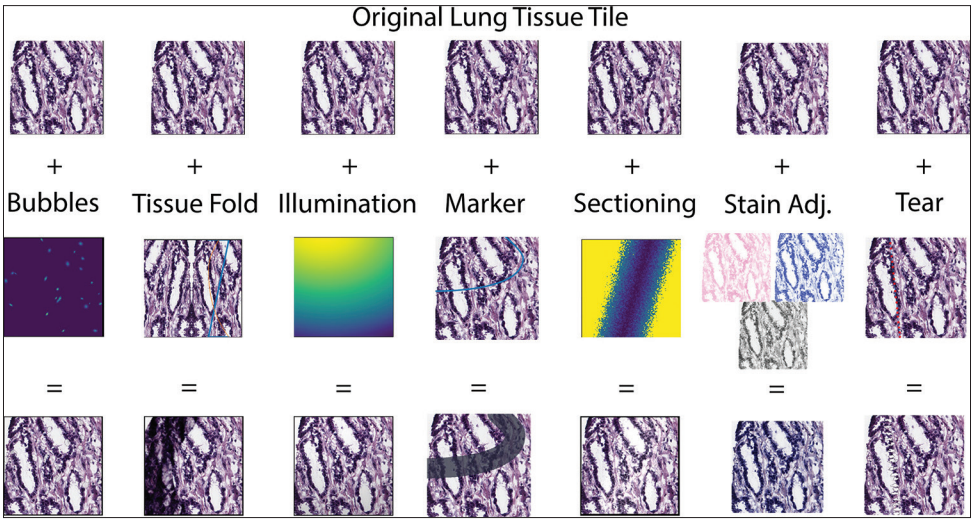

Figure 1: An example of seven types of simulated artifact, bubbles, tissue folds, uneven illumination, pen marks, sectioning artifacts, altered staining, and tissue tears. These artifacts are applied to the lung tissue tiles like in this example

Gaussian map is used to change only the brightness (value) of the image in HSV space.

Marker line

Pathologists often use marking pens to delineate regions of interest on a histology slide, for measurement, or other uses. While a pathologist would recognize that these structures were not a feature of the tissue sample, an algorithm may not be able to accurately distinguish the two. As such, a marker artifact was generated to test algorithms for their sensitivity to this kind of change. The marker artifact has a fixed width and follows a spline path from one edge to another, with a single intermediate knot between the start points and endpoints. The color is randomly generated from a range of dark colors, and the marker has an alpha of 0.75, giving it a slightly transparent appearance. Within the generator itself, more of these parameters are tunable, though for the purposes of this study, the default settings were used.

Sectioning

Sectioning artifacts are the result of unevenly cut regions of the tissue section during microtomy leading to varying thicknesses, and thus varying staining, across the slide.[15] This was simulated by generating a randomized line with a fixed relatively large width, and a slightly randomized edge. Within this sectioned region, to whiten the image, the saturation of the image is decreased, and the brightness is increased by half that amount. The amount of saturation increase is randomized within a range, and the effect is stronger near the middle of the section than the outsides of the sectioned region. The result is to make a region that looks thinner than the rest of the unaltered image.

Stain alteration

Even with a relatively consistent and well-known stain pattern like hematoxylin and eosin (H and E), there can be variability in the stain concentration and application can vary.[23] The stain alteration artifact involves artificially changing the staining concentrations of an H and E image. This is performed using the stain deconvolution GitHub package to determine the relative concentrations of H and E, and background using a generalized Ruifrok-Johnston color deconvolution.[24,25] Each of the stain levels is then adjusted up to 1.25× to 3× their original levels. Alternatively, the levels can be adjusted down from between 80% and ⅓% of their original levels. The level of change is chosen at random, and independently for H and E. The background can be increased to 1.5× original or decreased to ⅔ of the original levels. Afterward, the image is reconstructed with the new stain concentrations, on a full slide level, rather than a random tile by tile basis in the same manner as other artifacts.

Tissue tears

Slides are made by cutting tissue blocks into very thin sections, typically 3–5 μm using a microtome; however, this process does not always work perfectly. Due to technical issues during the cutting process, for example, when hard particles are present in tissue samples or due to vibrations in the knife blade, tears can occur within the cut tissue section. While these tears can have multiple patterns, dependent on the cause of the artifact, the generated tears were designed to mimic Venetian-blind pattern tears.[16] While the tears generated by this algorithm are configurable, they appear in the range of chattering tears to a continuous tear.

Tears are generated by first creating a randomized two-point spline or a line, which the tear will follow. The tear can start or end away from the edge (up to 15% of the size of the image), though this will only happen 50% of the time for each end of the tear. Once the tear path is generated, the center of each chatter in the tear is generated 20–40 pixels apart at random.

The tear pattern is created because of the layered nature of the point generation and can be tuned to have more perpendicular or more inline spread. The first layer has only 3–8 points, and 50% of the total spread, so it lays out the general shape of the tear. The second and third layers have their number of points generated based on the density of points in the end map, to

create the rest of the fill of the torn region. A smoothing layer and edge randomization is applied after that to make the tear look more natural. This tear shape then has a 2-pixel edge that is colored a darkened version of the average pixel color. The rest of the tear is colored background white, and the entire tear layer is overlaid on top of the image with a slight alpha transparency.

Experimental design

Artifacts were generated on both datasets and evaluated with the previously described models. The first dataset was Lung Cancer subtype classification by Coudray *et al*. where the tile inputs were used to call whether a tissue sample was adenocarcinoma, squamous cell carcinoma, or normal. The second pathology dataset was an internal kidney tissue dataset, where the goal was to perform image segmentation and classify pixels into six potential labels. While these models and datasets clearly differ in goals, they show the variety of applications that can be evaluated for their robustness to synthetic artifacts.

For both the datasets, there were 18 different levels of manipulations performed across the 7 different artifacts, with different percentages of tiles evaluated [as described in Table 1]. The varying percentages of artifacts were meant to provide a reasonable approximation of how often these artifacts might occur on a sample. To evaluate the raw effect of the artifact, each of the generated artifacts was also applied to all the tiles in the 100% experiments. In addition, a control study was run to evaluate the performance of the models on unperturbed data.

Due to the large number of samples and experimental conditions involved, we utilized the Snakemake workflow engine[26] to reproducibly execute the full prediction and image manipulation pipeline. The pipeline takes in a list of studies to apply, as well as percentage of tiles to manipulate for each study. Tiles were pseudorandomly selected independent of location within a slide using a random number generator seeded by the slide and experiment to compare different experiments on the same slides. The pipeline then proceeds according to the workflow described in Coudray *et al*., except for intercepting the tiled images to apply the relevant image manipulations [Figure 2]. While the kidney segmentation was built in MATLAB, the same artifact generation workflow was applicable.

The cancer genome atlas lung adenocarcinoma/lung squamous cell carcinoma/normal classification

The Coudray model takes in a pathology slide from a biopsy of a patient suspected to have nonsmall cell lung cancer, and classifies the slide as normal tissue, adenocarcinoma, or squamous cell carcinoma. The original authors trained the model on tissue slides from the cancer genome atlas (TCGA) lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) projects.[27,28] TCGA is an invaluable collection of molecular and phenotypic data from thousands of cancer patients across over 33 cancer types all collected under a rigorous protocol. All TCGA data that do not identify the source patient are available publicly, so we used the same TCGA dataset used to train the model to make our evaluations.

Table 1: A description of the 18 experiments performed based on the seven artifacts: bubbles, tissue fold, uneven illumination, marker line, uneven sectioning, stain alteration, and tissue tear. An additional control experiment was performed on the unaltered tiles to evaluate the baseline performance of the models

Artifact Type	Percent of Altered Tiles in Experiment
Bubbles	20%
	100%
Tissue Fold	20%
	100%
Uneven Illumination	10%
	50%
	100%
Marker Line	50%
	90%
	100%
Uneven Sectioning	15%
	30%
	100%
Stain Alteration	10%
	30%
	100%
Tissue Tear	15%
	100%
Control	0%

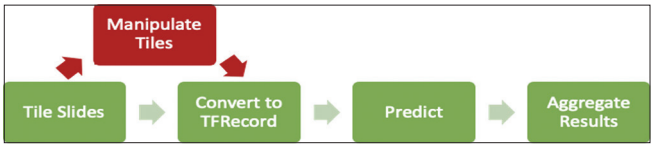

Figure 2: Schematic of processing steps undertaken by our Snakemake workflow. The "manipulate" tiles step in red was only applied to experimental studies

We downloaded the entire corpus of tumor and normal tissue slides from the National Cancer Institute's Genomic Data Commons Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive/) for the TCGA LUAD[27] and TCGA LUSC[28] studies. Tissue slides were identified by the presence of "01" in the sample barcodes, while normal tissue was identified by the presence of "11" in the barcode per the TCGA barcode documentation. We obtained pretrained weights, and preprocessing, prediction, and tile score aggregation scripts from the original model authors.

Kidney pathology internal dataset

An internal histopathology dataset of trichrome-stained kidney tissue samples was used to evaluate the effects of artifact addition on a deep learning model used for segmentation rather than classification. Instead of a whole tile or slide classification, this model used the DeepLab V3+ network based on the ResNet18 architecture to provide a label to each

of the tile's pixels.[29] Each pixel was labeled as one of the six possible classes (listed in decreasing prevalence): Tubules, interstitium, open glomeruli, arterioles, miscellaneous, and globally sclerosed glomeruli. The miscellaneous label was assigned to the remaining areas, including small black spot artifacts from the staining process. These labels were provided by an experienced pathologist for the training set that was used to train the deep learning model.

The kidney tissue slide images were preprocessed by splitting them into 256 × 256 RGB images. These images were also stained normalized, using Reinhard normalization from the Stain Normalisation Toolbox.[30] Data augmentation including reflection, rotation, and translation was included to increase the diversity of training examples. This model architecture was implemented using MATLAB software. A held aside set of 36 tissue samples or 4744 tiles was used as a testing set and where the artifacts were applied to evaluate the robustness of this trained model.

Results

Lung tissue classification results

To better compare the two datasets, these results will be focused on analyzing performance at the individual tile level. These examples display substantial variance as each prediction is made on a 512 × 512 slice of the image, with no aggregation of more and less informative regions. Stain alteration, tissue folds, marker lines, and sectioning artifacts show some of the largest variances in probabilities [Figure 3]. However, while the median probability did decrease in some of the more extreme examples, most of the probability change was increased variance. As would be expected, the shift in probability was usually related to the percentage of tiles that were altered within an experiment.

The subclass area under the receiver operating characteristic (AUROC) of the lung tissue classification model shows a significant but small decrease compared to the unaltered control for many of the artifact types [Figure 4]. Reflecting the pattern seen in probability variance, the artifact types with the greatest decrease in AUROC were stain alteration, tissue folds, marker lines, and uneven sectioning.

Kidney pathology results

The kidney pathology segmentation model had the same set of artifact experiments performed, and then, the AUROC was calculated on a per tissue component type (subclass) basis [Figure 5]. The artifacts which had the largest effect on the segmentation accuracy and AUROC were tissue folds, stain alteration, and marker lines.

Ten tiles were collected in more detail to investigate the changes in segmentation patterns in more depth. Two examples are provided in Figures 6 and 7, one which had multiple subclasses (four of the six potential subclasses) and another which consisted of the most common two subclasses (tubules and interstitium). Of these two examples, the initial segmentation on the tile with four subclasses was worse than the two-subclass tile. That four-subclass tile was also more affected by artifacts in general than the two-subclass tile.

Even in these few examples, while often the areas altered by artifacts have the most change, the changed labels are not always limited to that area. For example, the addition of a tissue tear through part of a region changed an entire area labeled as tubules to open glomeruli. Interestingly, marker and tissue fold artifacts sometimes produce the "miscellaneous" label which represents areas that had been labeled as an artifact in the initial sample by a pathologist. Overall, as might be expected from a deep learning model, the relationship between artifacts added and labels changed is nonlinear.

Discussion

We were able to successfully create artifacts for histopathology slides to apply to these tiles based on the literature. These artifacts come in a variety of forms and could be applied at varying levels of frequency. We were able to use these artifacts to stress test this machine learning model for lung cancer tissue. While there was a dose–response effect when increasing the presence of artifacts and increased uncertainty in these models, the effect level varied between the segmentation and classification tasks.

An important limitation of the lung cancer dataset is that because we worked off a pretrained model, we did not have information about which slides in the TCGA dataset were used as training examples. This makes it quite likely that many slides we included in our evaluations were already seen in training, and as such may not be a fair evaluation of the model's performance. Especially in the cases of experiments in which not all tiles were manipulated, it is quite possible that the model had already learned from that example, grossly inflating the prediction probability. While the reductions in probability we did see gives us confidence in our approach, future efforts should attempt to find an independent classification dataset for use in all studies.

The uncertainty surrounding the training and testing split was not an issue in the kidney pathology. As such, any changes in performance were on the held aside testing dataset. The changes in subclass AUROC were somewhat larger for the segmentation dataset but more variable across the subclasses. One interesting finding was the appearance of a miscellaneous artifact label in some of the marker and tissue fold experiments. There may be some value in having building in artifact classes into a segmentation model to increase the detectability of artifacts in a pipeline. One other note to make is that the slides had been stain normalized as part of the preprocessing for kidney tissue segmentation where the lung tissue classification samples had not been. This may mean that the segmentation model as part of a pipeline would have a reduced susceptibility to stain artifacts.

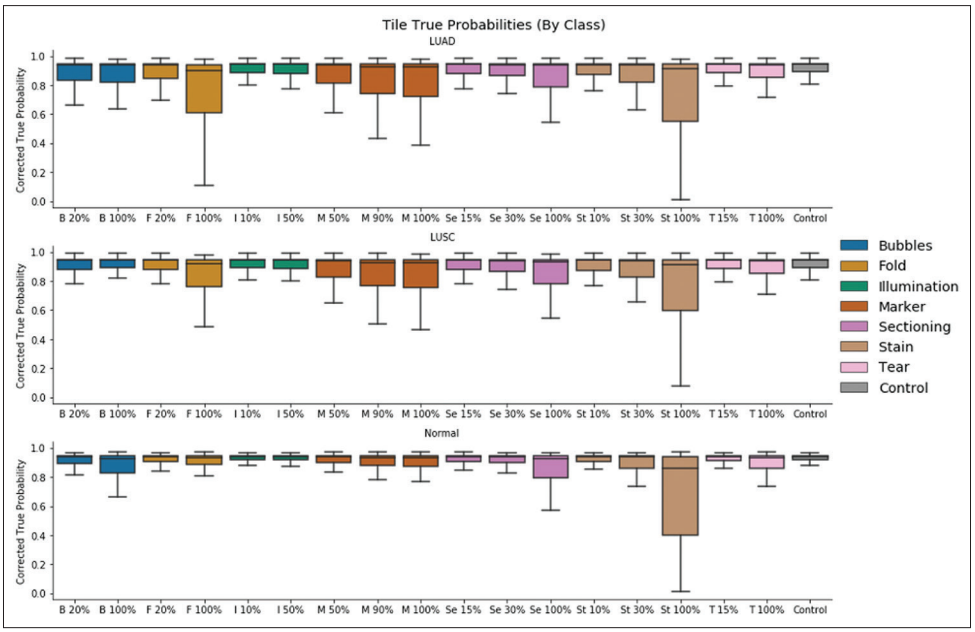

Figure 3: Tile level average predicted probabilities after artifacts were added. The three tissue types were split into separate categories, and the average probability is shown relative to the control set of images. In some cases, the probabilities had more spread, indicating some tiles had higher uncertainty

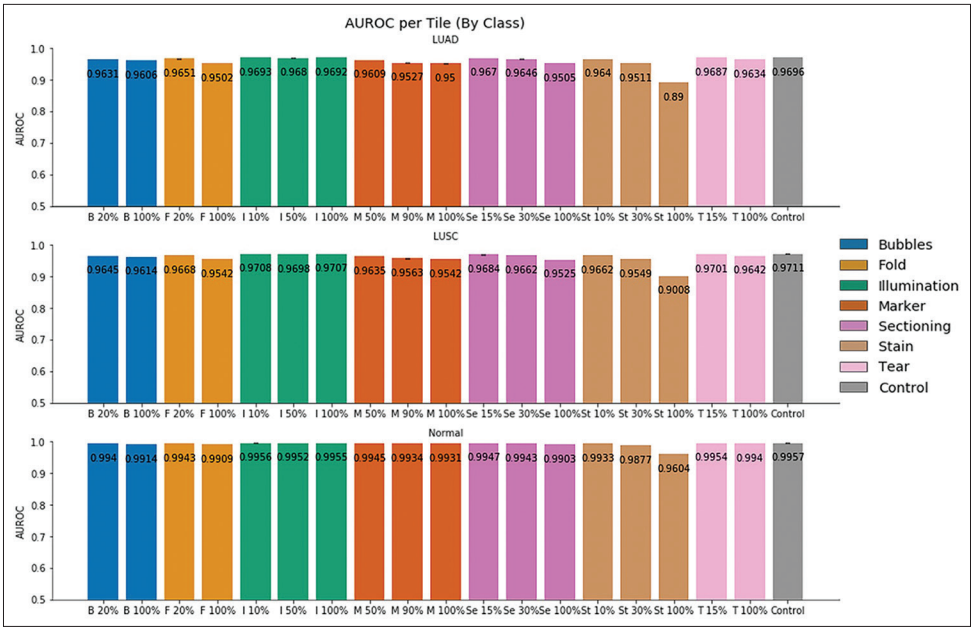

Figure 4: Tile level area under the receiver operating characteristics with confidence intervals after artifacts were added. The three tissue types were split into separate categories, and the subclass area under the receiver operating characteristic is shown relative to the control set of images. On a tile level, predictive performance was somewhat decreased by the artifacts introduced. Note that the area under the receiver operating characteristic y-axis ranges from 0.50 to 1

Future work

While seven artifact types have been included in this version of this testing, there are a large number of other potential artifact candidates that could be simulated. We consulted with pathologists (A.M and J.H) to determine some of the most common problems that occur, but many more artifacts exist. For example in frozen samples, artifacts like ice crystals or tissue damage can occur that are specific to the mode of tissue preservation.[31] In nonpolar solvents or paraffin embedding of fatty tissues, some of the fat can dissolve and leave voids behind.[17] In addition, the parameterization of these generated artifacts could be tuned and evaluated for its effects. Finally, there was no bias between the classes in the artifacts generated, as normal, LUAD, and LUSC were treated identically, but this may not necessarily be the case in a worst-case scenario. Overall, this framework is

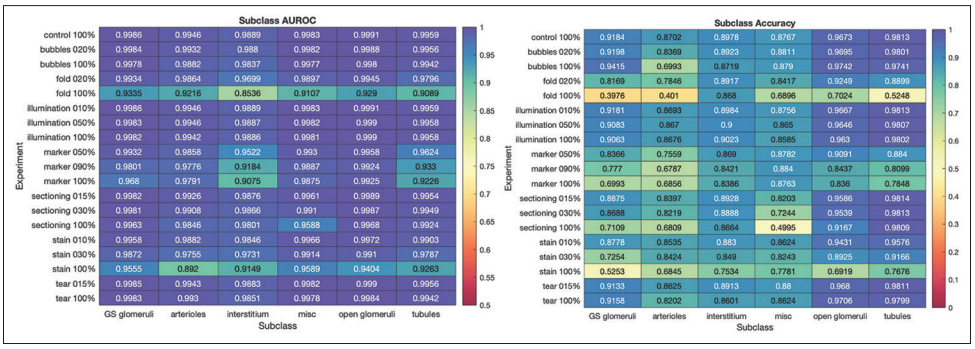

Figure 5: Area under the receiver operating characteristic and accuracy of kidney segmentation model after addition of artifacts, broken down by tissue component type (subclass). Area under the receiver operating characteristic focuses on the change in probability score produced by the model and its effect, whereas accuracy shows the change in predicted class. The baseline performance in the control experiment is described at the top of each chart. While tissue fold, marker, and stain alterations show the biggest changes, there is a lot of variability between subclasses. The prevalence of each subclass in the ground truth varies by several orders of magnitude across the six subclasses

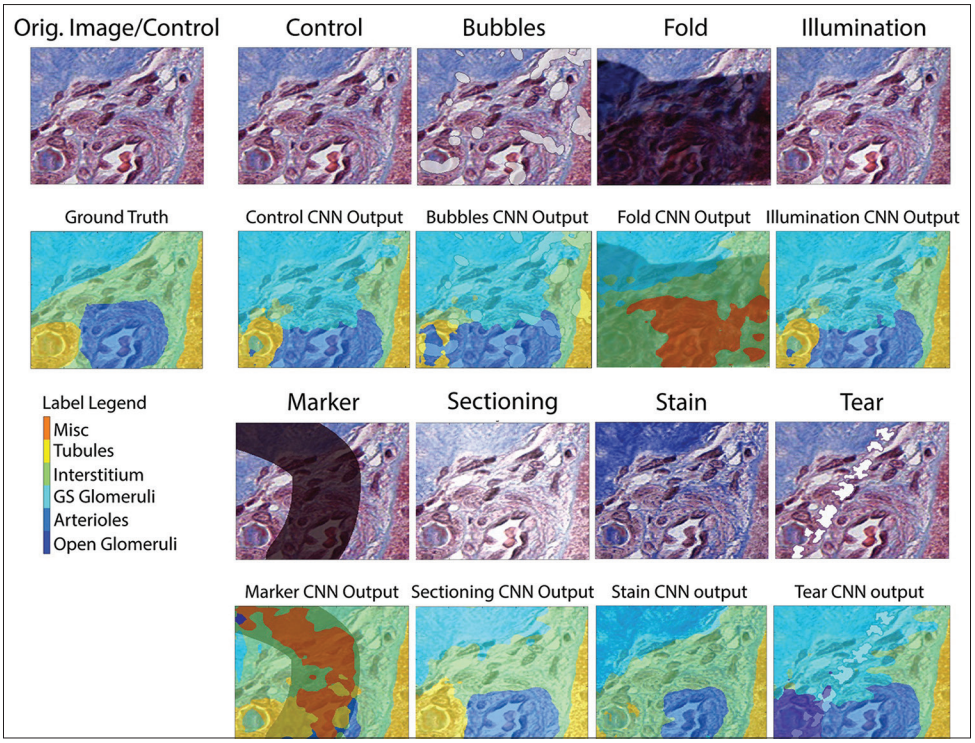

Figure 6: A selected example of the effects of artifacts on a single kidney tissue sample tile. This tile is notable for the presence of four tissue component types (subclasses) in the ground truth. Notably, the baseline model had difficulty properly identifying interstitium in this tile. Tissue folds and marker line artifacts both resulted in the "miscellaneous (Misc)" label originally designated for stain deposit artifacts

meant to be flexible and additive so that artifacts that occur frequently can be evaluated and simulated using pertinent generative models.

Using a GPU-equipped desktop allowed for the testing of a large number of tiles and a variety of generated artifacts. Processing the entire set of histopathology slides under the array of imaging artifacts introduced took approximately 6 days of computational time. However, to further tune the parameterization of artifacts introduced, more computational power would likely be necessary. This artifact generation approach is highly parallelizable and would also lend itself well to a distributed computing framework. Even still, further code optimizations are worth investigating.

We also believe this same framework and toolset could be applied to additional digital pathology models. Of specific interest would be models that look more at global patterns within a tissue slide. Our current approach randomly manipulates tiles without any regard for a spatial relationship, as our tested model does not account for tile relationships. However, we can imagine scenarios in which a fold through the entire slide causes effects more global than a single tile but does not affect the entire image. Adapting these methods for models that can account for such structures might identify more global failure modes in those models.

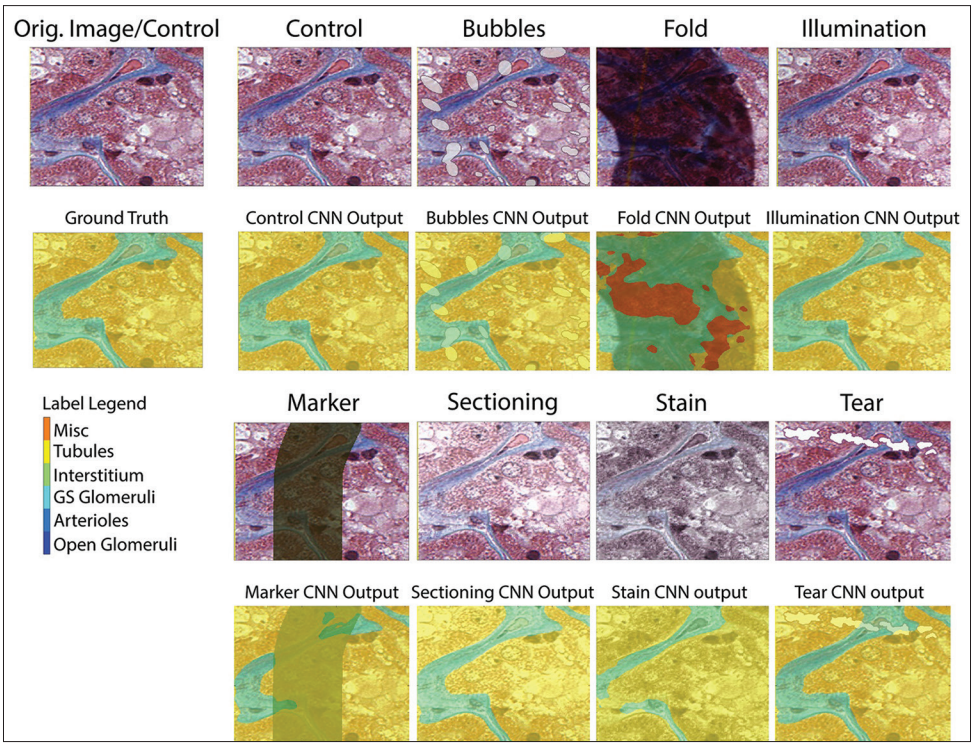

Figure 7: A second selected example of the effects of the impacts of artifacts. This example had only two tissue component types (subclasses) represented in the ground truth label, tubules, and interstitium. Of the ten tiles selected for further inspection, eight of ten had only these two most common labels. While tissue fold and marker artifacts did have the most effect on the labels, overall most of the artifacts applied had limited impact on this tile.

Conclusions

Based on the experiments conducted here, this package can help show where complex histopathology models may fail due to artifacts. It also helps to highlight the importance of preprocessing pipelines that can identify and handle artifacts. Several of the artifacts with the largest effect are those that change the color and/or contrast of the image (i.e., stain, sectioning, marker, and folds). As staining is known to vary across sites, it further emphasizes the need for consistent stain normalization. Those areas with artifacts can be dramatically altered, and the effect is often nonlinear and model and task dependent. Artifact generation can be a useful tool to increase the trustworthiness of models and provide a testing framework to understand the failure modes of deep learning models.

Availability: Code documenting our entire workflow, as well as input dataset manifests, is available at https://github. com/Systems-Imaging-Bioinformatics-Lab/histopath_failure_modes.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, *et al.* Clinical applications of machine learning algorithms: Beyond the black box. BMJ 2019;364:l886.
2. Shamout F, Zhu T, Clifton L, Briggs J, Prytherch D, Meredith P, *et al.* Early warning score adjusted for age to predict the composite outcome of mortality, cardiac arrest or unplanned intensive care unit admission using observational vital-sign data: A multicentre development and validation. BMJ Open 2019;9:e033301.
3. Jennings L, Deerlin VM, Gulley ML. Recommended principles and practices for validating clinical molecular pathology tests. Arch Pathol Lab Med 2009;133:13.
4. McPherson RA. Henry's Clinical Diagnosis and Management by Laboratory Methods: First South Asia Edition_e-Book. India: Elsevier Health Sciences; 2017.
5. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019;322:2377-8.
6. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28:231-7.
7. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, *et al.* Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine. 2019;25:1337-40. doi:10.1038/s41591-019-0548-6.
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.
9. Adam Bonislawski. Digital Pathology Poised to Take Off With FDA Clearances, AI Applications. 360Dx. Available from: https://www.360dx.com/cancer/digital-pathology-poised-take-fda-clearances-ai-applications. [Last accessed on 2019 Dec 12].
10. Paige Obtains CE Marks for AI-Based Breast Cancer Detection, Prostate Cancer Grading Tools. 360Dx. Available from: https://www.360dx.com/regulatory-news-fda-approvals/paige-obtains-ce-marks-ai-based-breast-cancer-detection-prostate. [Last accessed on 2020 Dec 14].
11. Hologic Obtains CE Mark for AI-Based Cervical Cancer Screening Platform. 360Dx. Available from: https://www.360dx.com/regulatory-news-fda-approvals/hologic-obtains-ce-mark-ai-based-cervical-cancer-screening-platform. [Last accessed on 2020 Dec 14].

12. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559-67.
13. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs]. 2015 Dec 11. Available from: http://arxiv.org/abs/1512.00567. [Last accessed 2019 Dec 12].
14. Lung Adenocarcinoma. LUNGevity Foundation. 2015 Jul 29. Available from: https://lungevity.org/for-patients-caregivers/lung-cancer-101/types-of-lung-cancer/lung-adenocarcinoma. [Last accessed on 2019 Dec 12].
15. Sampias C. H&E Basics Part 4: Troubleshooting H&E. 2018 Nov 15. Available from: https://www.leicabiosystems.com/knowledge-pathway/he-basics-part-4-troubleshooting-he/. [Last accessed 2019 Dec 03].
16. Taqi SA, Sami SA, Sami LB, Zaki SA. A review of artifacts in histopathology. J Oral Maxillofac Pathol 2018;22:279.
17. Chatterjee, S. Artefacts in histopathology. J Oral Maxillofac Pathol 2014;18:S111-6.
18. Beghi E, Beretta S, Carone D, Zanchi C, Bianchi E, Pirovano M, *et al.* Prognostic patterns and predictors in epilepsy: A multicentre study (PRO-LONG). J Neurol Neurosurg Psychiatry 2019;90:1276-85.
19. Kothari S, Phan JH, Stokes TH, Wang MD. Pathology imaging informatics for quantitative analysis of whole-slide images. J Am Med Inform Assoc 2013;20:1099-108.
20. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: An open-source quality control tool for digital pathology slides. JCO Clin Cancer Inform 2019;3:1-7.
21. Mowat A. Commentary to: The urinary microbiome in patients with refractory urge incontinence and recurrent urinary tract infection: (Zhuoran Chen, Minh-Duy Phan, Lucy J Bates, Kate M Peters, Chinmoy Mukerjee, Kate H Moore, Mark Schembri). Int Urogynecol J 2018;29:1783.
22. Popham EJ, Abdillahi M, Vickers H. The biting mechanism of the tsetse fly-a reappraisal of a film made by the late Professor R. M. Gordon, Dr. W. Crewe and Mr. J. Brady, using the scanning electron microscope. Trans R Soc Trop Med Hyg 1973;67:25-6.
23. Babak Ehteshami Bejnordi, Nadya Timofeeva, Irene Otte-Höller, Nico Karssemeijer, Jeroen A. W. M. van der Laak. Quantitative analysis of stain variability in histology slides and an algorithm for standardization. Vol. 9041. 2014. Available from: https://doi.org/10.1117/12.2043683. doi:10.1117/12.2043683.
24. Ruifrok AC. Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol 2001;23:291-9.
25. grfrederic/deconvolution. GitHub. Available from: Available from: https://github.com/grfrederic/deconvolution. [Last accessed on 2020 May 05].
26. Köster J, Rahmann S. Snakemake–A scalable bioinformatics workflow engine. Bioinformatics 2012;28:2520-2.
27. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, *et al*. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543-50. doi:10.1038/nature13385.
28. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519-25.
29. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv:1802.02611 [cs]. 2018 Aug 22. Available from: http://arxiv.org/abs/1802.02611. [Last accessed 2020 Nov 17].
30. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Comput Graph Appl 2001;21:34-41.
31. Desciak E, Maloney M. Artifacts in frozen section preparation. Dermatol Surg 2000;26:500-4.